2007 IIB 4B7 Dr D M Holburn
dmh@eng.cam.ac.uk

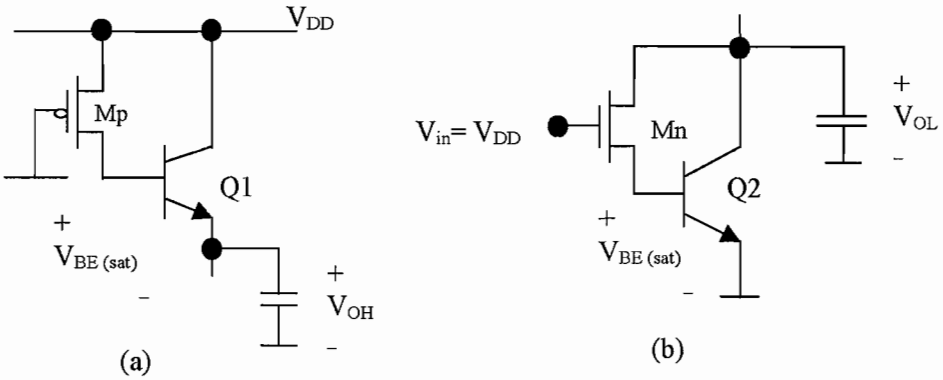
2007 Part IIB Module 4B7 VLSI Design, Technology and CAD

Answer:

(a) DC analysis.

Consider first V_{in}= 0 V. M_p is on, while M1 and Mn are off. Since Mp and M1 form an inverter, the base of Q1 is high at a voltage of $V_{DD} - V_{onMp}$ and it is in the on-state ; M2 is on and grounds the base of Q2 driving it off. The output high voltage V_{OH} for this case can be calculated from the subcircuit shown in figure (a). Noting that Q1 will eventually enter saturation, we have $V_{OH} = V_{DD} - V_{BE(sat)}$ (neglecting the voltage drop on Mp when Mp is fully on).

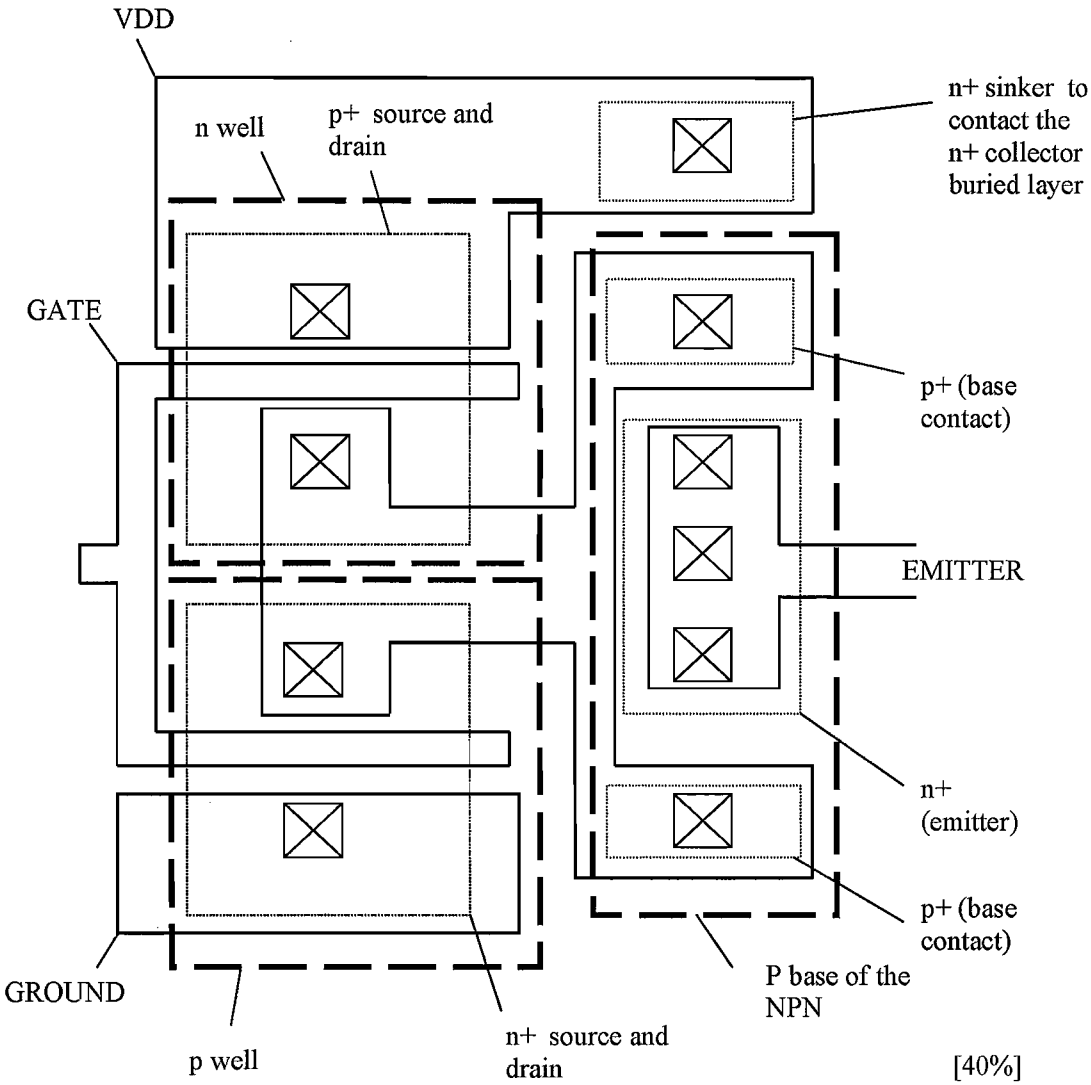
The subcircuit for the case where $V_{in} = V_{DD}$ is shown in figure (b). Now we see that Mp is off while M1 and Mn are on. M1 connects the base of Q1 to the ground, driving it off. This in turn shuts off M2 so that Q2 is on. The output low voltage V_{OL} is seen to be $V_{OL} = V_{BE(sat)}$ since Q2 induces a base-emitter drop (the analysis is neglecting the voltage drop across Mn when Mn is fully on). The drawback of this configuration is that the output logic swing is reduced from V_{DD} by $2V_{BE(sat)}$.



- M1 and M2 are used to provide paths to remove charge from the base of the BJT transistors thus increasing the gate switching speed.
- For loads with high capacitances, the Bi-CMOS NOT gate is much faster than the standard CMOS inverter gate. The drawback of the Bi-CMOS NOT gate configuration is that the output logic swing is reduced from V_{DD} by $2V_{BE(sat)}$.

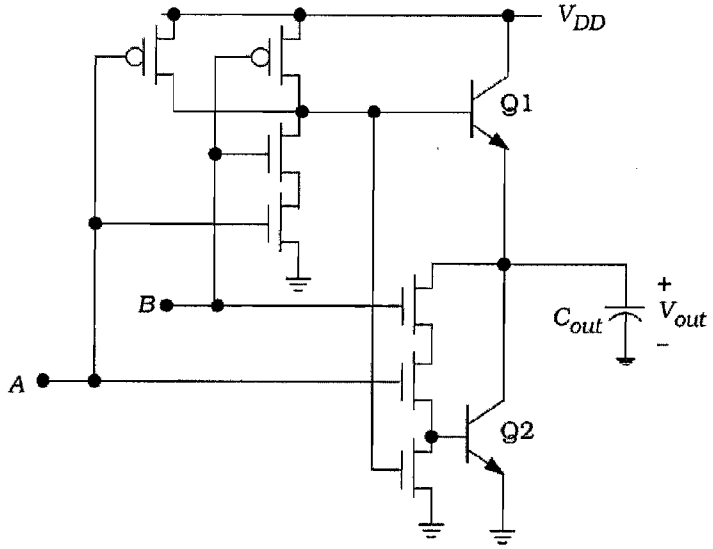
[30%]

(b) The layout is based on an inverter with the output connected to the base of an NPN transistor.


VDD
n well
p+ source and drain
n+ sinker to contact the n+ collector buried layer
GATE
p+ (base contact)
EMITTER
n+ (emitter)
p+ (base contact)
GROUND
P base of the NPN
p well
n+ source and drain

[40%]

c)


V_{DD}
Q1
B
A
C_{out}
+
V_{out}
–
Q2

[30%]

2. a) Advantages:

- excellent electrical isolation between devices or blocks of devices (less leakage, no latch-up)
- less area consumed (no buried layers)
- (alternatively one can mention the high junction temperature)

Disadvantages:

- self-heating (the buried oxide layer acts as a thermal barrier)
- expensive (the SOI wafer can be 5-10 times more expensive than a standard bulk silicon wafer)

[20%]

b)

Wafer bonding

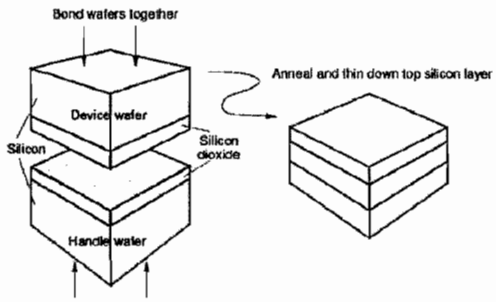
The basic technique relies on the fact that polished and flat wafers, when brought into contact at room temperature, are attracted to each other by van der Waals forces and ``bond". To strengthen the bond between the two wafers, a post-bonding anneal at high temperature is usually performed, and the top silicon wafer is then polished to create a thin silicon-on-insulator layer suitable for device manufacturing

Unibond -SmartCut

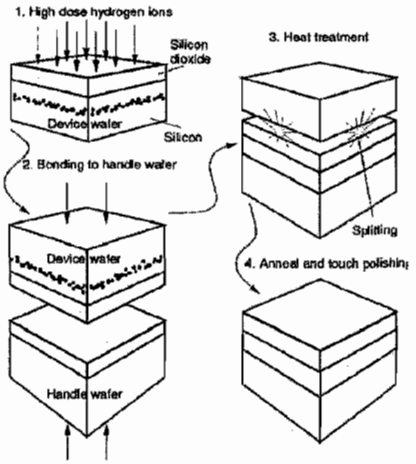
The device wafer, which has a layer of silicon dioxide on top of it, is implanted with a high dose of hydrogen ions (between 3.5 $x10^{16}$ and $1x10^{17}$ cm^{-2}), after which it is bonded to the handle wafer. A heat treatment at 600 C divides the wafers along the line of the implanted hydrogen, leaving behind a thin and uniform silicon-on-insulator layer on the handle wafer which requires only a final high-temperature anneal and touch polish to yield the finished wafer.

Wafer bonding is much cheaper than SmartCut. Smart cut however is much better in defining accurately the thickness of the SOI layer.

Wafer bonding



Unibond-SmartCut SOI technology



[30%]

c) (i) Time dependent dielectric breakdown (TTDB)

Good quality thermal oxide films have dielectric breakdown strength of 10 MV/cm or more.
However, oxide film failure over time even in lower electric-field intensity (conditions of practical use) is a major cause of failure. Gate oxides are generally affected by this. When an electric field applied to an oxide film causes the injection of holes into the oxide film to occur and it consequently causes traps to be made in the oxide film. As the number of traps increases, an electric current via the traps is due to hopping or tunneling. If the number of traps continues to increase and the traps connect between the high voltage and low voltage terminal, the connection carries a high current that causes the gate oxide film to break down.

(ii) Hot carrier Injection

It is generated in MOSFET by the large channel electric fields near the Drain region. Mechanism: Carriers (electrons or holes) that flow into the high electric field area are accelerated by the strong field and gain substantial energy. Some of the carriers have enough energy (that is to say they are hot) to overcome the electric potential barrier existing between the Si substrate and gate oxide film. These hot carriers are injected and subsequently trapped into the gate oxide film. They form a space charge or inversion layer and over a period of time they can affect the threshold voltage (Vth), transconductance (gm) or breakdown.

[20%]

(d) The acceleration factor is defined as the ratio between the failure time in use and the failure time in the acceleration test.

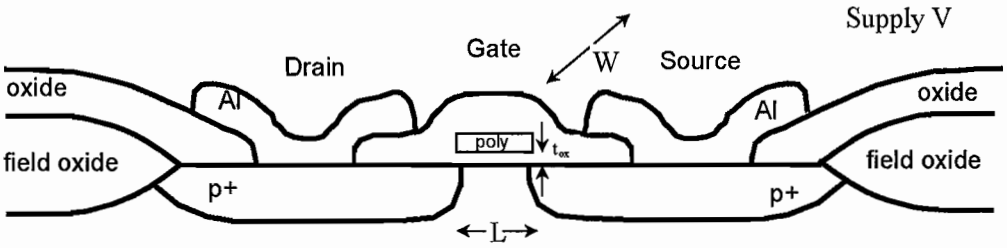
$AF = tf_{use} / tf_{test}$.

$AF = [Juse/Jtest]^{-2}\, e(Ea/kTuse) / e(Ea/kTtest) = [Juse/Jtest]^{-2}\, e(Ea/k)\,(1/Tuse - 1/Ttest) = e(0.8 / 8.6x\ 10\text{-}5\ )\,(1/(273+40) - 1/(273+125)) = 25 \times 571 = 14275$

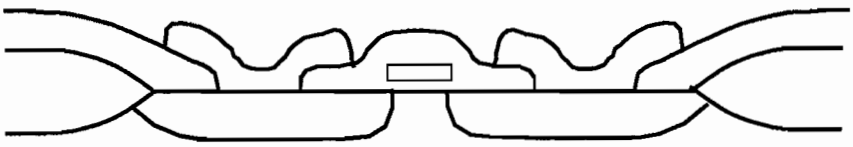
Average failure in the field of use= $(6 \cdot 10^{-5}/h) / 14275 = 4.2 \cdot 10^{-9}/h = 4$ FIT (approx)

3 In *constant field* scaling, geometric dimensions are modified (typically reduced) and electrode voltages are also scaled in order to maintain electric fields constant. Historically, device dimensions were scaled from about 6 microns to 1 micron without change in Vdd (*constant voltage scaling*). This offered better delay reduction as well as cost reduction; it maintained continuity in supply and logic level standards. However, it has proved impracticable to push dimensions to sub-micron with Vdd unchanged, since the increasing electric fields impact the operation of the MOSFET, requiring other major process alterations (e.g. doping densities, etc). There would also otherwise be greater risk of breakdown. Some adjustments to other process parameters may also be required.

Assume all dimensions & voltages are reduced by a scaling factor k ~ 1, as in the diagram. The MOS transistor has critical dimensions L and W (channel length and width respectively) and gate oxide thickness t_{ox}. The applied voltages are represented by V. We shall consider scaling down all dimensions and voltages by a scaling factor k.



Scale by factor k to give L/k, W/k, t_{ox}/k, V/k



[20%]

Packing density	α	$1/LW$ increased by k^2	
Field at channel	α	V/t_{ox} unchanged	

Since carrier velocity is μE and distance travelled α L

Transit time τ thru channel	α	L^2/V decreased by k	

Hence carrier speed is increased by k

Capacitance at gate etc	α	LW/t_{ox}	decreased by k
Current I consumed I	α	CV/τ	decreased by k
DC Power consumption	α	IV	decreased by k^2

For interconnect, scaled in length l, width w, and thickness d, and dielectric thickness h, the key parameters are changed as follows:

Resistance	α	l/dw	increased by k
Capacitance to substr	α	lw/h	decreased by k
Current density J	α	I/dw	increased by k

Note that this assumes device currents are scaled down by k as above.

Time constant RC	α	RC	unchanged

Hence the speed of propagation of signals along interconnect is unchanged.

For linear circuit elements, there are additional considerations. While in digital design typically the smallest possible devices should be used to minimise parasitics and provide best speed-power product. This desire has driven the 'push' to smaller geometries in the microprocessor and memory industry.

Off-chip loads Devices may have to source or sink large off-chip loads, which may be resistive, capacitive, inductive, or any combination. They must therefore be sized larger to provide these currents. As for digital designs, operation at high frequencies normally dictates the use of small devices, since these have lower parasitics.

Power dissipation Linear circuits may operate at a range of source-drain voltages and currents. This may mean that power dissipation may be significant within devices: Maximum Power Transfer Theorem shows that in a CMOS output stage, power dissipation is maximised when the output terminal lies midway between the supply rails. The heat produced has to be absorbed and conducted away if destructive temperature rises are to be avoided, and calls for devices of larger W and L (though with W/L unchanged) to maintain the energy density at a safe value.

Balanced and matched devices Differential amplifiers and operational amplifiers require balanced and matched devices to minimise offsets and secure acceptable Common Mode Rejection Ration (CMRR). Because of process tolerances and statistical variations, this is less easily achieved with smaller devices.

Electrical noise Electrical noise may be an important factor in linear circuit performance. The principal sources in MOS circuits are: (a) thermal noise, (b) flicker noise. Of these, flicker noise may be most troublesome and it dominates at low frequencies, but may still be significant at frequencies in excess of 1 MHz. It can be reduced by keeping the device gate capacitance large (high $W \times L$); and by using large channel length L, although this reduces the available gain. The front-end stages of a low-noise amplifer need to be of large area, and should operate at low current levels. P-type devices tend to generate less noise than n-type.

Control of channel conductance Where transistors are used as active resistors, control of the resistance can be effected by use of a suitable choice of aspect ratio L/W, as well as by control of V_{GS}. In current sources and current mirrors, device aspect ratio may routinely be optimised to obtain current outputs in the desired ratios.

Major benefits – scaled devices allow:

- higher packing density
- greater speed of operation
- lower current consumption

The industry typically scales process generations with $k \sim \sqrt{2}$, which is roughly the ratio described in the question. The reduction in *Vdd* is consistent with constant-field scaling. This doubles the number of transistors per unit area with each generation and doubles transistor performance every two generations under constant field scaling. Process shrinks of $k \sim 1.05$ are commonly applied as a process becomes mature to boost the speed of components in that process

Problem areas

- Charge stored in transistor gate reduced by k^2, hence scaled devices (e.g. memories are more liable to soft errors
- Roff/Ron decreases as dimensions decrease and the role of sub-threshold conduction becomes more important. Hence static power consumption will become a more serious issue.
- Faster clock speeds - these have risen far faster than classical scaling would predict – with Vdd still somewhat higher (viewed historically) than constant field scaling would demand, have led to skyrocketing power dissipation.
- Dynamic power dissipation cannot continue to increase unchecked because it will be uneconomic to cool the chips.
- Increasing clock speeds allied with trend towards larger devices leads to longer interconnect delays as a function of τclock. Clock skews – differential timing changes – may rise as k^3. Manufacturers may attempt to circumvent this by use of Cu interconnect, low-permittivity dielectrics, and multiple interconnect layers scaled less aggressively
- Contact resistances rise as contact structures are scaled
- Some structures e.g. I/O pads, power amplifiers, do not scale
- Reduction in yield at smaller geometries
- Digital cells are typically well characterised in a new process before linear designs (amplifiers, oscillators, mixers) can be adequately verified and reliable models developed. This may delay the introduction of a new process for mixed signal applications.
- Increased fab. costs and lower yields at the beginning of process lifetime mean that for an evolving design that is sensitive to market price, the point at which transition should be made to a smaller process must be carefully judged.

[80%]

4 The account of design rules should include the following major points:

- Design rules allow a ready translation from circuit concepts to actual geometry in silicon
- They are the effective interface between the circuit/system designer and the fabrication/process engineer. They provide a reliable and workable compromise which is friendly to both sides.
- The designer is concerned to achieve:
  - best possible electrical performance - speed. noise margin, etc
  - minimum area of Si per circuit – lower cost, better yield, reliability
- The process engineer seeks:
  - to maximise tolerances on all parts – easier fabrication, better yield
- There are 3 basic tolerances that set limits to the shapes the designer can use
  - dimensional resolution governed e.g. by λ of light used in photolith, photoresist characteristics
  - alignment errors – registration, temperature changes, bowing/distortion
  - reproducibility of processing – wet etching, layer thickness control
- For practical purposes all 3 effects can be reduced to linear dimensions on a plan view of the mask layout. The permissible dimensions are often highly specific to a manufacturer's process.
- The simplest rules originate from the need for continuity and avoidance of unintended short-circuits. Layers such as polysilicon, metal and diffusion are associated with minimum dimensions and minimum separations. They may also be associated with ohmic resistance (electrical origin). Violation of these rules may lead (as in PCB technology) to open-circuits in conducting traces, or short-circuits, where tracks are too close.
- With metal Al interconnect it is necessary to ensure that the current density does not exceed about 10^9 Am^{-2}, otherwise there is risk of electromigration induced by transfer of momentum from the electronic carriers to metal atoms and causes progressive thinning of interconnect at circuit bottlenecks – e.g. as metal crosses a step. Interconnect width is hence governed by the anticipated peak (rather than mean) current, and not simply by lithographic considerations.
- Since fabrication involves several sequentially masked steps, there is a need to accommodate the possibility of mis-registration between successive masks. For this reason,
  - implant masks overlap the diffusions to which they correspond by a significant margin
  - polySi gates extend beyond the edge of the underlying diffusion
  - metal, diffusion and polySi are required to surround contact cuts by a significant margin
- It is possible to define an 'alignment tree' which summarises the statistical probability of mis-registration between related mask layers.
- The use of metal (Al/Cu) rather than polySi is dictated for
  - power distribution
  - signal transmission over significant distances – for example, clock lines, to avoid skew.
- Where significant currents are transmitted from one metal layer to transistors, or to another metal layer, the contact structure must be capable of carrying the current. Since contact conductance is proportional to cut perimeter (not area),

this is achieved through use of many minimum geometry cuts filling the available space.

- Other rules that may be mentioned: contacts and vias of fixed size, antenna rules, well/substrate tap spacings.

[60%]

Numerical Part

The units comprises 16,000 memory cells each driving 80 fF capacitance. Each is clocked at 40 MHz. In the worst case, a pattern of 0-1-0-1 etc on successive clocks will generate maximum dynamic current in these cells.

Whenever a cell output switches 0-1 or 1-0 a packet of energy 1/2 $CVdd^2$ is transferred between the supply rails for that stage. We assume that the dynamic dissipation arising from this dominates other effects. Note that if all inputs remain at 1 (or 0), every stage in the 2,000 bit register is presumed to stay in the corresponding state and no dynamic dissipation would be observed. This assumes no resistive or other losses of charge occur requiring that charge be replenished at each output (e.g. refresh).

In the worst case, each stage of the unit alternates its output state at each successive clock edges, giving rise to maximum dynamic dissipation. This will occur when each unit receives at its input a 0101010 waveform synchronised with the clock, and at half the clock frequency.

Each stage thus dissipates energy at a rate:

$\frac{1}{2} CVdd^2 \times fc$ where fc is the clock frequency

Hence the total power dissipation is thus

$W = 8 \times 2000 \times 40 \times 10^6 \times \frac{1}{2} \times 30 \times 10^{-15} \times 3.3^2 = 0.104$ W

Hence the average current consumption

$I = 0.104/3.3$ A $= 31.7$ mA [20%]

Let the interconnect width be W. Then the current density J is:

$J = 31.7 \times 10^{-3} / (W \times 0.5 \times 10^{-6})$ A m^{-2}

This must be significantly less than the electromigration limit of $\sim 10^{10}$ Am^{-2}. A factor of 10 is usually considered adequate. To satisfy this we have:

$J < 10^9$, giving:

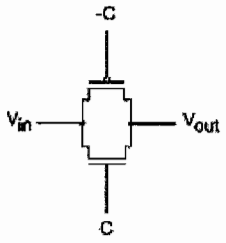
$$J = \frac{31.7 \times 10^{-3}}{W \times 0.5 \times 10^{-6}} < 10^9 \quad \text{which gives } W > \frac{31.7 \times 10^{-3}}{0.5 \times 10^{-6} \times 10^9} = 63.4\ \mu\text{m}$$

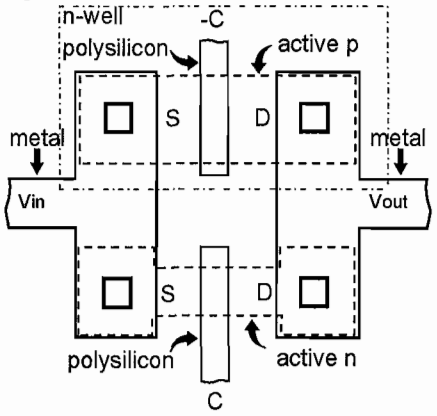
The total capacitance being driven is approximately 16 × 30 pF or 480 pF. To this needs to be added the capacitance being driven at the 8 output pins, which may each drive a capacitance of order 30pF, making a total of a further 240 pF. Hence the total true worst-case current may be twice that calculated. Typically the pads have their own suitably dimensioned power ring, but nonetheless in a conservative design verification of the total current will be necessary.

Note also that the peak current may be many times I, with current transients synchronised to clock edges.

[20%]

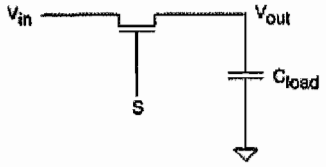
5 (a) A transmission gate in CMOS consists of a pair of complementary transistors with source and drain regions strapped together.

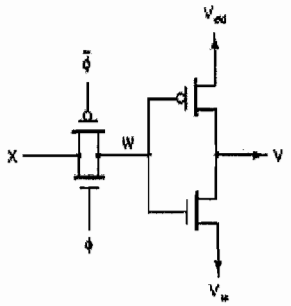




Note: well and substrate taps not shown

The two gate electrodes are driven with complementary control signals C and –C. When C is high and –C is low, both p and n channel devices are conductive. In the opposite situation, both are non-conductive. Note that the device is bilateral when seen from Vin or Vout. This allows its use in linear circuits.





D-type bistable

To understand the performance issue consider a single n-channel pass transistor used as a switch. The n- device is conductive when S is high (Vdd), non-conductive when S is low (Vss=0v). To allow a channel to form, V_{GS} must exceed Vt. If C is at Vdd then if Vin is also driven to a high potential around Vdd, Vout cannot rise above Vdd–Vt, typically a 1 volt drop. Thus if Vin were high, Vout would be a so-called weak low. Note that in low state at Vin is transferred reliably to Vout when S is high, since both Vin and Vo << V_{GS}-Vt. As a result it is impracticable to connect single transistor switches in cascade. Conversely, a p-type transistor can exert a strong 'high' but only a weak 'low'. Its control input is of course inverted cf. that for the n-device. By combining the complementary devices in parallel, a switch can be made which suffers from neither of these shortcomings.

In digital circuits T-gates may be used to realise multiplexers, which may be bilateral. They are commonly used to control feedback paths in sequential (memory) circuits. A major application is in the implementation of a dynamic D-type bistable –see above. Charge is stored on the parasitic capacitance at W.

Advantages
- o low device count for multiplexers and bistables
- o bilateral characteristic
- o high performance
- o can be cascaded

Disadvantages
- o effectively a passive device, does not repower logic levels
- o requires complementary control signals (extra logic)
- o may be sensitive to clock dispersion or skew

In analogue circuits T-gates may also be used in bilateral switches/Muxs for linear signals. A common usage is in sample-hold circuits or electronic exchanges.

Advantages
- o efficient switch with low offset voltage
- o good frequency response
- o good ratio Roff/Ron
- o compact structure

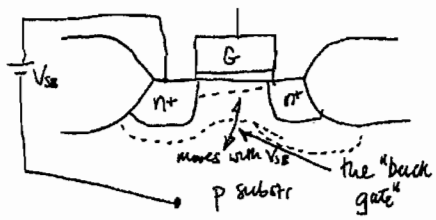
Disadvantages
- o Insertion loss may be significant and varies with Vin, Vout [50%]

5 (b) An MOS transistor consists electrically of charge stored in the dielectric layers, in the surface/surface states and in the substrate (or well) itself. Switching an enhancement-mode MOST from **off** to **on** consists of applying a gate potential to neutralises these charges and to cause the underlying semiconductor to undergo an inversion due to the E-field from the gate. Hence the threshold gate voltage can be written:

$$Vt = \phi_g + \frac{Q_B - Q_{SS}}{C_O} + 2\phi_{fN}$$

Here, ϕ_g is the WF between gate and Si (typically very small)
ϕ_{fN} is the Fermi potential between the inverted surface and the bulk silicon
C_O is the capacitance per unit gate area
Q_{SS} is the charge density at the $Si:SiO_2$ interface in the channel
Q_B is the charge in the depletion region beneath the gate oxide

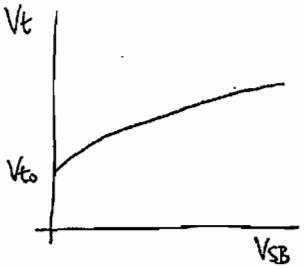
With the exception of Q_B, these are dependent only on physical/material parameters and process parameters. However, Q_B depends on ϕ_{fN} and the potential between the transistor source and the substrate, V_{SB}. This is the so-called *body effect*. It is also referred to as *back-gating*, since the substrate in effect acts as a form of gate situated 'behind' the channel.
Increasing V_{SB} causes the channel charge to be depleted; the perceived effect is that Vt is raised for a single transistor, according to the following:



Change in Vt is given by:

$$V_t = V_{t0} + \gamma\sqrt{V_{SB}}$$

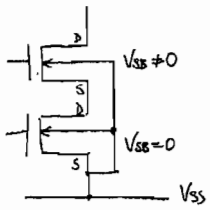
for nMOS devices, where V_{t0} is the threshold voltage for VSB = 0, and γ is typically 0.5 to 1.5, being process-dependent.



Where transistors are connected in series as in 2/3/4 … input static logic gates, hand computation of the transfer function becomes very difficult, owing to the body effect. The Spice simulator can model this effect accurately

The upper transistor has a higher Vt than the lower owing to body effect. This means that for multi-input gates, the switching level (Vdd/2 for the ideal inverter, is raised for NAND gates and lowered (NOR gates).

This has the subsidiary effect of eroding the noise margins in a corresponding way.



In fact, the switching level and noise margins will change according to which input/s change in a transition.

As far as transient response is concerned, transistors exposed to significant body effect will have a lower apparent conductance in the ON-state, assuming a fixed V_{DS}. As a result, multi-input gates will exhibit slower rise/fall times when parasitic capacitances are charged/discharged through series-connected transistors subject to body effect.

The designer can compensate for this effect by selecting devices of greater W/L in proportion to the lower conductance in the ON-state of affected devices. Such compensation would have to be done in the light of detailed simulation.

[50%]